A rule-based approach to unknown word recognition in Arabic

Lynne Cahill
NLTG, University of Brighton
Lewes Rd, Brighton
BN2 4GJ, UK
L.Cahill@brighton.ac.uk

Abstract

This paper describes a small experiment to test a rule-based approach to unknown word recognition in Arabic. The morphological complexity of Arabic presents its challenges to a variety of NLP applications, but it can also be viewed as an advantage, if we can tap into the complex linguistic knowledge associated with these complex forms. In particular, the derived forms of verbs can be analysed and an educated guess at the likely meaning of a derived form can be predicted, based on the meaning of a known form and the relationship between the known form and the unknown one. The performance of the approach is tested on the NEMLAR Written Arabic Corpus.

1 Introduction

The Semitic languages, especially Arabic, are linguistically interesting for a number of reasons, and are attracting more and more attention for both linguistic and socio-political reasons. One of the aspects of Arabic that makes it particularly interesting to linguists, namely the morphological complexity, is at once both appealing and the source of potential practical problems. It is appealing to linguists, for whom it offers interesting challenges in their descriptive frameworks, but for builders of NLP applications, it represents a significant challenge. In this paper, we are particularly interested in the derivational aspects of the morphology, whereby verb stems are derived from triliteral roots in well defined formal ways, and with varying degrees of regularity in the meanings of those derived forms.

Another aspect of the Arabic language that makes it both interesting and challenging is the fact that it is not actually a single language. There are many varieties of Arabic, with rather different status. Classical Arabic (CA) is the language of the Koran, and the historical ancestor of the other varieties. Modern Standard Arabic (MSA) is the modern version of CA and is, broadly speaking, the universal (i.e. not regional) standard variety of Arabic. Until recently, CA and MSA were the only varieties that were written – other, regional, varieties were only spoken. The situation is rapidly changing, with electronic communication increasingly involving written versions of the regional varieties. Even in traditional written forms, such as news reports, the vocabulary used in different geographical regions is different. For example, Khoja (2001) found that the percentage of out of vocabulary items in news reports from Egypt and Qatar was around double that found in Saudi news reports, Saudi Arabic being much closer to MSA than the other two regional varieties. Ways in which the present approach may assist in this problem will be discussed later.

The approach we describe here depends on a hierarchically organised lexicon, based on the DATR lexical representation language (Evans and Gazdar, 1996). The PolyLex lexical framework (Cahill and Gazdar, 1999) was developed originally with languages like English, German and Dutch in mind, but has been shown to lend itself to the description of Arabic templatic morphology (Cahill, 2007, 2010). The inheritance of information by default in this framework is fundamental to the approach we describe.

The problem to which we seek a solution is not one unique to Arabic. Any NLP system which wants to

Proceedings of the Twelfth Meeting of the Special Interest Group on Computational Morphology and Phonology (SIGMORPHON2012), pages 35–41,
Montréal, Canada, June 7, 2012.

process naturally occurring text will always have to deal to some degree with the problem of unknown or out of vocabulary (OOV) items. Whether these items are neologisms, errors or names, they need to be handled in some way. Solutions to this particular problem are unlikely to have a large statistical impact on the success rates of the processing applications, but that does not mean that they are not worth finding. While it is undoubtedly the case that many applications will work perfectly well with a word recognition rate of, say, 95%, supported by statistical approaches which provide syntactic information, there are other applications for which full semantic interpretation is desirable, if not necessary. It is such applications that the current paper addresses. We are only addressing a part of the problem, as this approach does not help recognise names or errors.

The particular approach described in this paper is based on the observation that a native speaker who encounters a word they have not seen before may, if that word is related to others that they do know, be able to make an educated guess at not only the syntactic category, but also the meaning of that word. To a large degree, that guesswork involves the specific context that the word occurs in, but native speakers will also have more abstract structural knowledge about their language which allows them to make guesses about words on the basis of their internal structure. For example, if an English speaker knows the word "confuse" and hears the word "confuser", even though they have most likely never before come across the latter, they will be able to at least guess that it means "someone/thing that confuses". Of course, with derivation the meaning relationship is not always transparent. So a person encountering the word "decider" for the first time may be surprised to find that it does not mean "one who decides" but rather a deciding match/game etc.. Such issues and other limitations of this approach will be discussed later.

2 Previous approaches

There has been a lot of work on how to handle OOV items, largely based on statistical approaches. Some are language independent (see e.g. Attia et al (2010), Adler et al (2008)) while others focus on specific languages (see e.g. Habash and Rambow (2005, 2007) and Marsi et al (2005) on Arabic and Adler and Elhadad (2006) on Hebrew, another Semitic language with similar morphological structure). The work by Habash and Rambow, for example, employs a form of morphological expansion to handle OOV items, but only makes use of the inflectional morphology of Arabic, not the derivational morphology as in the current approach.

Other approaches to morphological analysis in Arabic include methods to deal with OOV items. For example, Beesley and Karttunen (2003), describe a two-level approach which includes a general method for guessing OOV words which could certainly apply to some degree to Arabic, but it would not be able to take into account the linguistic (specifically semantic) information which is at the heart of the present approach.

3 PolyLex/PolyOrth

The PolyLex project (Cahill and Gazdar, 1999) developed multilingual lexicons of the morphology and phonology of English, German and Dutch, implemented in the lexical representation language DATR (Evans and Gazdar, 1996) which allows for default inheritance. Therefore, aspects of these languages that were shared could be inherited by default by each language.

In addition to the aspects of inter- and intra-language default inheritance, the other aspect of the PolyLex framework which contributes to the unknown word processing proposed here is the use of phonological structures, specifically syllables, to define morphological structures and relationships. Thus, in PolyLex, the lexical entries consist of specifications of the phonological forms of the syllable constituents (onset, peak and coda). These can be determined by morpho-syntactic features. For example, the English word *man* has default values for the onset (/m/), peak (/æ/) and coda (/n/), but a further value for the peak in the plural (/ɛ/). This is represented in DATR as[1]:

```
<phn syl1 onset> == m
<phn syl1 peak> == {
<phn syl1 coda> == n
<phn syl1 peak plur> == E.
```

The PolyOrth project (Cahill et al. 2006) further developed the representation so that orthographic

[1] In the DATR code, the SAMPA machine readable alphabet (Wells, 1989) is used.

forms are derived by means of a combination of phoneme-grapheme mappings and spelling rules. Both types of information include phonological and morphological determinants, so that, for example, the default mapping for any particular phoneme will depend on both its phonological position (is it in the onset or coda?) and on its morphological position (is it in a stem or an affix?). Both types of information are defined by means of Finite State Transducers (FSTs)[2]. This framework has been implemented and tested on English, German and Dutch, and now extended to Arabic (Cahill, 2010). The Arabic lexicon allows for forms to be defined in Arabic script, Roman transliteration or phonological representation.

4 Arabic verbal morphology

The Arabic languages have around 280 million speakers. They belong to the Semitic language family, and share many linguistic features with other Semitic languages, such as Hebrew and Maltese. Much work in both theoretical and computational linguistics has focused on the so-called templatic morphology of the Semitic languages.

The key area of Arabic morphology addressed in this paper is the verbal derivation. Verbs in Arabic are typically based on a tri-literal root, consisting of three consonants. Inflectional variation involves interdigitating these consonants with vowels which indicate the tense, aspect and mood. In addition, the three consonants can be differently arranged (doubled, swapped etc.) to form distinct Forms (or measures, also known as binyanim[3], especially when applied to Hebrew). These are essentially derivations and form distinct verbs with different meanings. For example, the tri-literal root *k-t-b* has the core meaning "write". The forms *katabtu* and *aktubtu*, represent the active perfective and active imperfective first person singular forms of "write", namely, "I wrote" and "I write". The second Form or measure verb *k-tt-b* also has the inflectional variations, but has the meaning "cause to write", thus the two actual forms *kattabtu* and *akttabtu* have the meanings "I caused (someone) to write" and "I cause (someone) to write" respectively.

There are fifteen different Forms in CA, but fewer in the modern varieties. In MSA there are ten that are commonly found, although two more are found rarely. The regional varieties all make use of fewer. While some of the Forms have clear transparent meanings, others have far less clear or apparently random meaning relations.

The following descriptions of the meanings of the ten Forms is adapted from Scheindlin (2007):

I. The basic Form – all verbs have this form. May be transitive or intransitive.
II. Almost always transitive. If a verb exists in both Form I and II then I will often be intransitive and II transitive (*write* (I) → *cause to write* (II)). If I is transitive then II may be ditransitive. II may also involve an intensifying of the meaning on I, e.g. *kill* (I) → *massacre* (II).
III. May involve reciprocity, e.g. *follow* (I) → *alternate* (III).
IV. Like II, mostly transitive, and often matched with intransitive in I.
V. Often involves a reflexive element, e.g. *know* (I) → *teach* (II) → *learn* (V).
VI. Like III, often involves reciprocity, e.g. *fight* (I) → *fight each other* (VI).
VII. Mostly reflexive, resultative or passive. Roots that are transitive in I are intransitive in VII. E.g. *break* (I) → *be broken* (VII).
VIII. Often reflexive for verbs that are transitive in I, e.g. *divide* (I) → *part* (VIII).
IX. Very restricted in application, only applying to verbs indicating colours and defects, e.g. *turn yellow*.
X. Often associated with asking for something associated with the Form I verb, e.g. *pardon* (I) → *apologise (ask for pardon)* (X).

As is clear from these descriptions, the meaning relationships are not fully predictable, but they can give some hints as to the likely meaning of an unknown verb. As the framework relies on default inheritance, the assumption that any definitions may be overridden by more specific information means that even very approximate definitions are still valuable.

[2] The PolyOrth project was inspired by Herring (2006). However, while Herring uses one-stage FSTs, the PolyOrth project used two levels of FST, including a separate treatment of "post-lexical" spelling rules.
[3] We will use the term "Form", capitalised to avoid confusion with the more usual use of "form".

5 Arabic in syllable-based morphology

A small sample lexicon of Arabic in the PolyLex framework is presented in Cahill (2007). What makes this account different from most accounts of the morphology of the Semitic languages is that it requires no special apparatus to allow for the definition of so-called "templatic" morphology, but makes use of the same kind of equations as are required for ablaut and consonant devoicing, for example, that are found in English, German and Dutch.

5.1 The default, Form I root

The main part of the account addresses a single verb root, namely *k.t.b*, 'write', and generates all possible Form stems for perfective, imperfective and participle, active and passive. The approach is based on defining the leaves of syllable-structure trees, with the consonants of the triliteral stems occupying the onset and coda positions, and the vowels (syllable peaks) being defined according to the morphosyntactic specification, as in the example of *man* above. To illustrate this, the figure below shows the default structure for a triliteral root, with no vowels specified. The default structure is a disyllabic root, with the first consonant occupying the onset of the first syllable, the second consonant occupying the onset of the second syllable and the third consonant occupying the coda of the second syllable[4].

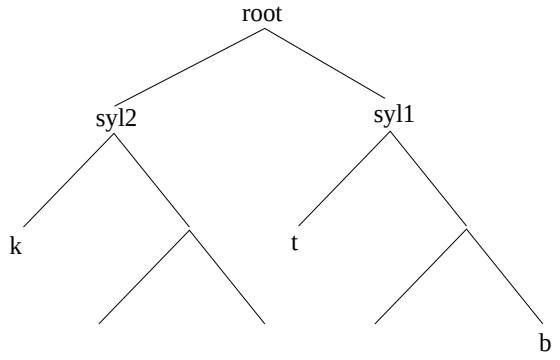

Figure 1: the structure of /katab/

[4] The syllable position is indicated by simple numbering. Syllables can be counted from either right of left. For languages which largely use suffixation, it makes more sense to count from the right, as for Arabic here.

5.2 The other Form stems

As described in Cahill (2007), the remaining nine forms have their default structure defined in similar terms. Figure 2 depicts the inheritance of forms from each other. This inheritance is for the syllable structure definitions, so the Form II structure is the same as the Form I structure except that the first coda has the value of the second root consonant, the same as the onset of the second syllable. The definitions are all incremental, so that each Form specification only supplies one or two pieces of information.

5.3 Meanings

The original lexicon was designed to demonstrate that the complex relationships between phonological, morphological and orthographic forms in Arabic could be captured in the PolyLex/PolyOrth architecture. There was no semantic information in the lexicons at all. For the present experiment, we have added very basic semantic information for the 100 verbs we have included. Most of these are Form I verbs, but there are some Form II, Form IV and Form V verbs. Where possible, we have represented the meanings of the verbs of Forms other than I in terms that can be generalised. For example, the verb *apologise* has the meaning expressed as ASK FOR PARDON[5].

The lexical hierarchy, in addition, defines a default meaning expression for each Form. For Form VIII, for example, this is:

```
<meaning> == ask for "<formI meaning>"
```

which says that the meaning is simply the string "ask for" followed by the meaning for Form I for the root[6].

5.4 The full lexicon

[5] For this small experiment, the exact representation of the meanings is not important. It is assumed that in a genuine application will have its representations which would be included in the lexicon, or for which a mapping can be defined.

[6] The quotes around the path <form1 meaning> indicate that it is to be evaluated at the original query node, i.e. the root node in DATR.

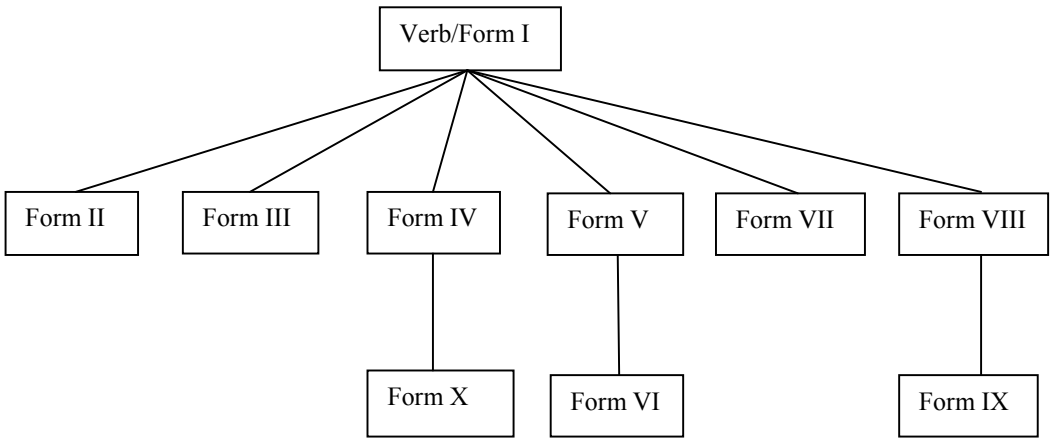

Figure 2: The inheritance of Forms

As stated above, the lexicon we are working from has only 100 verbs. There are no verb roots for which we have more than one Form. This is a very small number, but for each verb in the lexicon there are a theoretically possible further nine verbs which may be derived from the same root. The lexicon will recognise any theoretically possible verb from the roots it knows about, although it does not have semantic information explicitly provided for a large proportion of these verbs.

6 Using the lexicon for word recognition

The highly structured, hierarchical lexicons are not designed to be used as they are within NLP applications. The information in them is cached in a lookup table which can be used for either generation or comprehension, with entries which look like this:

كتب	k-t-b	katab	stem	p, a	k-t-b	I	write
كتتب	k-tt-b	kattab	stem	p, p	k-t-b	II	[cause to write]

The first column is the form in Arabic script, the second is the transliteration, the third is one possible full (vowelised) form, the fourth and fifth give the morphological analysis, the sixth is the triliteral root it is derived from, the seventh is the Form and the last is the translation. The first row, which has the Form I entry, has a translation which was provided explicitly in the lexicon but the second gets its meaning by default. This is indicated by the square brackets. In use in an application, these meanings would be used more cautiously, possibly in conjunction with other methods, especially making use of context.

The lookup table often provides more than one possible entry for a single form, especially when the form is unvowelised.

6.1 Testing

In order to test the approach, we tested the recognition of all verbs in the NEMLAR written corpus (Attiyya et al., 2005). The corpus provides versions with POS tagging, which enabled us to extract each verb. There were a total of just over 40,000 forms tagged as verbs, approximately 11,000 of them unique forms. Initial tests only took those forms which were tagged as having neither prefix nor suffix, a total of 1274 verb forms[7]. These included forms which were inflectionally distinct, and once these forms were collapsed, the total number of verb forms is 577. Of these, 32 occurred in our initial lexicon of 100 verbs.

These tests showed that of the remaining 545 unknown verbs, 84 could apparently be analysed as derived forms of one of our existing verbs. This

[7] The decision to use only those forms without prefix of suffix was simply made to make the testing process simpler and to ensure that the results were not skewed by the presence of consonants in prefixes or suffixes.

was determined by checking the main entries in an online Arabic dictionary and comparing the meanings given to those generated by the lexicon. This was a very promising figure, given the very small size of the lexicon.[8]

In the next testing phase we looked more closely at these forms. There are two ways in which the analyses may not be appropriate. The analysis might not be an appropriate (or at least not the most appropriate) one. This is not a major problem since we are dealing with a situation in which we frequently have multiple possible analyses for a word, so generating a number of possibilities from which an application must choose is exactly what is required. The second issue is the question of whether the meanings generated are useful. In order to check this we manually compared the generated meanings against the actual meanings for a sample of the verbs in question. We found that just over half of the verbs we checked had meanings which were at least clearly connected to the generated meaning. For example, the stem علّم (*teach*) is clearly related to the stem علم (*know*), and turns out to be the second Form ("cause to X") of the root for which *know* is the first Form.

6.2 Analysis of results

The verbs for which meanings were generated fit into three broad categories. First there are verbs for which the derived Form appears in dictionaries with the same meaning as that for Form I, possibly as one of its meanings. Thus, for example, the Form VIII verb *ktatab* had the meaning "wrote", the same as the Form I *katab*. There were 23 verbs in our set of 84 for which this was the case.

The second category consists of verbs for which the meaning is related in the way suggested by our earlier analysis. 22 of the verbs came into this category.[9]

Finally, the last category consists of verbs whose meaning is not related in the way suggested. This is the most problematic class, and unfortunately the largest in the small test set we are working with. However, in most, indeed nearly all, of these cases, the generated meaning was not wildly different from that in the dictionary. Closer inspection suggests that simply improving the meaning relations, and allowing more than one additional possible lexicon entry for some Forms would improve the performance significantly.

7 Discussion and conclusion

This paper has described a small experiment to test a novel rule-based approach to unknown word recognition in Arabic. Although testing is at an early stage, the initial results are promising.

The experiment described is intended to address a small part of the overall problem of unknown words. In some respects it can be viewed as more of a technique for extending an existing lexicon than for dealing with OOV items at runtime. However, it would be possible to enable an application to have access to the default lexical information at runtime, to allow this.

Another area in which the above technique may prove particularly useful is in the processing of regional varieties of Arabic. As stated above, Khoja (2001) found that even texts apparently written in MSA were twice as likely to have unknown words in texts from Egypt and Qatar than from Saudi Arabia. This suggests some variation in the vocabulary, most likely involving "leakage" of vocabulary items from Egyptian and Qatari Arabic into the MSA used by those speakers. As the morphological patterns of derived verbs are different in the different regional varieties, taking these patterns into account will provide further possible interpretations. The PolyLex structure allows the definition of similarities and differences between the lexicons of languages and dialects that are closely related.

7.1 Limitations and future work

The experiment described here is a very small scale one, and the lexicon is extremely small. The representation of meaning is also extremely simplified. It is possible that the approach described simply could not be scaled up to a size useful for an application. However, there is a range of ways

[8] There were some difficulties with transliteration which mean that these figures may not be fully accurate.

[9] This is clearly a case of subjective judgement, and from a non-native speaker these judgements may not be accurate.

of representing meaning, including linking to an external ontology, which could also be implemented in the lexicon described.

The next phase of work is to fully evaluate the results of the initial tests, followed by further more extensive testing. It is envisaged that an iterated cycle of testing and extension of the lexicon could lead to a lexicon large enough to be useful and robust enough to handle significant (if still small) numbers of OOV items.

Subsequently, and further down the line, development of a lexicon (or lexicons) for the vocabulary of regional varieties, linked to the MSA lexicon in the PolyLex framework will help to exploit the similarities. That is, the lexicon for, say, Egyptian Arabic assumes that, by default, words are the same as in MSA, with only those words (morphemes, phonemes etc.) which differ requiring specification.

Acknowledgements

The work described here was partly supported by the ESRC (Economic and Social Research Council, UK) as part of the project: **RES-000-22-3868** *Orthography, phonology and morphology in the Arabic lexicon*. We are grateful to the anonymous reviewers for their helpful comments.

References

Adler, Meni and Michael Elhadad. () An Unsupervised Morpheme-Based HMM for Hebrew Morphological Disambiguation. *COLING-ACL 2006*, pp. 665-672.

Adler, Meni, Yoav Goldberg, David Gabay and Michael Elhadad. (2008) Unsupervised Lexicon-Based Resolution of Unknown Words for Full Morphological Analysis. *ACL-08 : HLT*, pp. 728-36.

Atiyya, Muhammed, Khalid Choukri and Mustafa Yaseen. (2005) *The NMELAR Written Corpus* ELDA.

Attia, Mohammed, Jennifer Foster, Deirdre Hogan, Joseph Le Roux, Lamia Tounsi and Josef van Genabith. (2010) Handling Unknown Words in Statistical Latent-Variable Parsing Models for Arabic, English and French. *NAACL HLT Workshop on Statistical Parsing of Morphologically Rich Languages.* pp. 67-75.

Beesley, Kenneth and Lauri Karttunen. (2003) *Finite State Morphology* Chicago : CSLI.

Cahill, Lynne. (2010) A Syllable-based Approach to verbal Morphology in Arabic. *Workshop on Semitic Languages, LREC2010*, Malta, 2010.

Cahill, Lynne. (2007) A Syllable-based Account of Arabic Morphology. In Abdelhadi Soudi, Antal van der Bosch and Günther Neumann (eds.) *Arabic Computational Morphology* Dordrecht : Springer. pp. 45-66.

Cahill, Lynne, Jon Herring and Carole Tiberius, "PolyOrth: Adding Orthography to a Phonological Inheritance Lexicon", *Fifth International Workshop on Writing Systems*, Nijmegen, Netherlands, October 2006 (available at http://www.nltg.brighton.ac.uk/projects/polyorth).

Cahill, Lynne and Gazdar, Gerald. (1999) The PolyLex architecture : multilingual lexicons for related languages. *Traitement Automatique des Langues*, 40 :2, pp. 5-23.

Evans, Roger and Gazdar, Gerald. (1996) DATR : a language for lexical knowledge representation. *Computational Linguistics*, 22 :2, pp. 167-216.

Habash, Nizar and Owen Rambow. (2007) Arabic Diacritization through Full Morphological Tagging. *NAACL HLT 2007*pp. 53-56.

Habash, Nizar and Owen Rambow. (2005) Arabic Tokenization, Part-of-Speech Tagging and Morphological Disambiguation in One Fell Swoop. *ACL 2005*, pp. 573-80.

Herring, J. (2006) *Orthography and the lexicon*, PhD dissertation, University of Brighton.

Khoja, Shereen. (2001) APT: Arabic Part-of-speech Tagger. *Proceedings of the Student Workshop at the Second Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL2001).*

Marsi, Erwin, Antal van den Bosch and Abdelhadi Soudi. (2005) Memory-based morphological analysis, generation and part-of-speech tagging of Arabic. *ACL Workshop on Computational Approaches to Semitic Languages.* pp. 1-8.

Scheindlin, Raymond P. (2007) *501 Arabic verbs* Haupage: Barron.

Wells, John. (1989) Computer-coded phonemic notation of individual languages of the European Community. *Journal of the International Phonetic Association*, 19 :1, pp. 31-54.